AUTOMATION OF THE DETECTION OF LUNG CANCER CELLS IN MINIMAL SAMPLES OF BRONCHIOALVEOLAR LAVAGE

Carlos Ortiz-de-Solorzano[1,2,3], Thomas Pengo[1,2], Miguel Galarraga[1], Arrate Muñoz-Barrutia[1,2]

[1]Oncology Division, Center for Applied Medical Research, University of Navarra, Pamplona, Spain
[2]Department of Electrical and Electronic Engineering, University of Navarra, San Sebastian, Spain
[3]Department of Histology and Pathology, University of Navarra, Pamplona, Spain

ABSTRACT

We present the hardware and software specification of a quantitative, multidimensional and multispectral microscopy system designed for the detection of lung cancer using minimal samples of bronchoalveolar lavage (BAL). BAL samples were stained using FICTION: Fluorescence Immunophenotyping and Interphase Cytogenetics as a Tool for the Investigation of Neoplasms. Our system allows preliminary immunophenotypic detection of rare cancerous candidate cells, followed by accurate three-dimensional analysis of genomic integrity, to confirm or refute the initial assessment. Our results show that our automated analysis can accurately assist a human expert in the diagnostic evaluation of BAL samples.

1. INTRODUCTION

Lung cancer is one of the most prevalent causes of death in Western countries. Increasing its low survival rate –mainly due to late diagnosis- requires detecting the disease when it is still at microscopic, pre-surgical stages. To this end there is a push towards the discovery of new biological markers of pre-or-early neoplasia in biological fluids. Bronchoalveolar lavage (BAL) is a novel method to obtain a spreads that contains secretions, cells, soluble proteins, lipids and other chemical constituents from the epithelial surface of the lower respiratory tract. In patients with lung cancer, BAL samples may also contain few cancer cells exfoliated from the surface of the tumor. Finding these rare –low probability- cancer cells in BAL samples requires highly specific labeling of the cells and accurate detection.

Our labeling and detection method works in two stages: first, candidate cancer cells are identified based on the expression of a cancer biomarker. Those cells are then screened for genomic aberrations, characteristic of solid tumors. This double labeling technique is known as Fluorescence Immunophenotyping and Interphase Cytogenetics as a Tool for the Investigation of Neoplasms (FICTION) [1]. FICTION combines immunophenotypic labeling of the cancer biomarker with multiple fluorescent in situ hybridization (MFISH) of DNA sequences. Manually searching for rare tumor cells in macrophage and debris plagued BAL samples assumes consistent evaluation of immunofluorescence positivity, followed by three-dimensional counting of multiple FISH targets. This requires long hours at the microscope in low ergonomic working environments. This is prone to serious inter and intra-observer variability, making automation highly recommended. In this paper we present the hardware and software specification of a mutidimensional, multispectral image acquisition and analysis platform for the analysis of FICTION samples.

2. INTEGRATION

Automating the analysis of FICTION samples involves automating three concurrent activities: image acquisition, analysis and storage. Unsupervised acquisition is available in most microscopy platforms. Software packages like Analysis, Metamorph, MATLAB or open-source alternatives like ImageJ or Octave can deal with non-too specific image analysis tasks. Some advanced image storage solutions exist, like OME and OMERO (www.openmicroscopy.org). However, no existing platform seamlessly integrates all three tasks. Our challenge was to integrate these three aspects to create a system able to perform efficient searching for cellular objects in fluorescence microscopy. Using a combination of available open source software and in-house written software, we have developed such a system. We now briefly describe some aspects of this integration task.

2.1. Hardware

Our hardware platform is composed of a fully automated microscope (Zeiss Axioplan2ie), a cooled-CCD monochrome camera (Photometrics CoolSnap cf by Roper Scientifics), an automated stage and slide loader by LUDL, a tunable LCD filter, the VariSpec; two multiprocessor workstations, linked by a Gigabit Ethernet connection and an RGB filter. The workstations both run RedHat Enterprise Linux.

2.2. Software Integration

To implement the integration requirements, the system was divided in four parts: three reflect the imaging functions and one represents the integration.

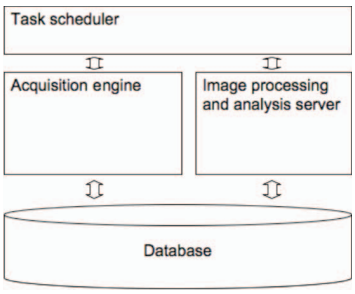

Figure 1. The simplified architecture, where each part corresponds to a different functionality of the system.

2.2.1. The acquisition engine
This part is responsible for the acquisition of the images, controlling the devices and storing the results in the database. It is connected to the database for image and metadata storage and receives requests from the task scheduler. The structure reflects a progressive abstraction from the hardware devices to the high-level operation layer. The lower layers deal with the communication with hardware devices, while the higher ones control operations that require complex interactions with the microscope. This progressive abstraction permits easy driver and device replacement, without having to modify the higher levels. The integration of the new driver is generated automatically by SWIG, open source software for wrapper generation [2].

5.2.2. The image processing and analysis server
This part is responsible of all image processing and analysis tasks. The images are retrieved from and the results stored in the database. To integrate these procedures, we use an open source middleware for component networking, called the Internet Communication Engine (ICE). ICE provides a set of libraries and protocols that enable the communication between software components. It currently supports the creation of servlets in different languages (C++, Java PHP, Python, Ruby, C#, Visual Basic), which means that the analysis routines can be written in any of these languages. The interface to the procedures is specified in the Slice language: the client library and the server stub are compiled to Java and C++ respectively. All image analysis tools were programmed as C extensions of the Diplib 3D image processing library (www.diplib.org).

2.2.3. The task scheduler
This module is responsible for the coordination of all acquisition and image processing tasks. Jobs are received from the user level, split into tasks and assigned to either the acquisition engine or the image processing and analysis server. The job scheduler is the main component, which enables the concurrent management of job queuing and execution. To this purpose, an open-source job scheduler is used, named Quartz (http://www.opensymphony.com /quartz), which has been used successfully both in the academic and in industry.

2.2.4. The data
Keeping track of the images resulting from the analysis is a critical role in any integrated imaging application. In our application, the information is progressively added to a hierarchical data structure, which has been tailor-written for this. In C, this is implemented as a structure with pointers, which results in a tree of objects. Each node contains the information from the analysis and keeps track of all the data resulting from the algorithms. The database layer is a relational database coupled with an object-relational mapping system written in Perl and remotely accessible through XML-RPC. The database system has been developed by the OME Consortium (http://www.openmicroscopy.org). The information from the hierarchical data structure can be easily accessed from DipLib and MATLAB.

3. IMAGE ACQUISITION AND ANALYSIS

3.1 Sample preparation
We used three training samples and two validation samples. One sample (T1) was a spread of cells from a lung cancer cell line (H460). The remaining four samples were BAL samples sprinkled with known amounts of H460 cancer cells. The samples contained 36 (T2), 71 (T3), 49 (V1) and 99 (V2) cancer cells, respectively. The base BAL material was obtained from a patient suffering from severe bronchial infection, but with no symptoms of lung cancer. All samples were stained using an antibody for the nuclear protein hnRNPA1, which is highly expressed in nearly all lung cancers but is expressed only at basal levels in normal lung epithelia. The samples were also labeled with the LaVysion kit, which contains four FISH probes that target three common loci of genetic alterations in lung cancer (5p15.2 SpectrumGreen, 8q24 SpectrumOrange, 7p12SpectrumRed), along with a centromeric probe (6c SpectrumAqua).

3.2 Image acquisition
The entire sample was scanned at low magnification (20X) Three spectral channels -Alexa 350 (B), SpectrumAqua (G) and SpectrumRed (R)- were acquired. All images were preprocessed and analyzed to detect immunopositive cells. Areas occupied by candidate cancer cells were then revisited and re-imaged as Z-stacks of images (40X) in all spectral channels. Nuclei in the stacks were analyzed for the detection and enumeration of FISH probes.

3.3 Image Analysis

3.3.1 *2D* Analysis

Preprocessing. Before the analysis, we corrected for the inhomogeneity of the light source and intensity fluctuations caused by uneven flurochrome distribution. Finally, short spectral shifts were eliminated by shifting the images a number of pixels precalculated on test images.

Detection and classification of nuclei. We applied an adaptive threshold [3] on the Alexa 350 channel to separate all nuclear objects from the background. This left us with all areas occupied by immunopositive cells, but also with numerous macrophages, and other cell types that showed some background fluorescence. Most macrophages and immunonegative nuclei are white because their fluorescent emission is very similar in the three channels. Immunopositive cells instead appear as blue-redish objects with medium to high chromatic purity. Highly saturated -blue- objects correspond to organic debris. It is therefore easy to reject white and very pure blue objects by fixed-thresholding the Hue (H) and Saturation (S) channels of the images. To that end, we converted the RGB image to HIS format and then fixed-thresholded the images using the hue (H) and saturation (S) channels. We avoided false negatives by forcing that no immunopositive cell be classified as immunonegative. Then we measured the morphology of all segmented objects to classified them as isolated nuclei or clusters of nuclei using Sequential Minimal Optimization-Support Vector Machine (SMO-SVM) [4]. We used a linear SVM classifier trained to avoid false positives (clusters classified as nuclei), to prevent incorrect spot counting.

3.3.2 *3D* Analysis

Preprocessing. Similar to what was described for 2D, the acquired Z-stacks were pre-processed to correct for inhomogeneities of the light, background fluctuations and chromatic shifts. Then, the images were deconvolved using a Maximum Likelihood Estimation algorithm provided by the Huygens (SVI, The Netherlands) software. Finally we used blind spectral unmixing to eliminate cross-talk between the spectral channels due to overlap of the spectra.

Segmentation of nuclei and FISH signals. Nuclear volumes were extracted from the background using an adaptive threshold applied to the immunofluorescence channel. Then all areas occupied by nuclei were analyzed in the corresponding FISH channels. FISH signals were segmented using a Top Hat algorithm with morphological reconstruction (Figure 2d-f).

4. RESULTS

4.1 Training of the classifiers

4.1.1. Detection of immunopositive cells

Thresholding T2 and T3 BAL training samples produced 9783 and 11573 objects respectively, 36 (T2) and 71 (T3) of which were immunopositive cancer cells. Based on the H and S threshold values, our algorithm detected all 36 (T2) and 71 (T3) immunopositive tumor cells, along with 60 (T2) and 32 (T3) objects with tumoral origin. Therefore the combined classification error was 0.43%. All errors corresponded to false positives (FP) since the threshold was purposely set to avoid false negative results (FN).

4.1.2. Classification of nuclei vs. clusters of nuclei

One hundred and two (102) objects segmented from sample T1 were used -51 isolated nuclei and 51 clusters of nucle-). All but 3 objects (isolated nuclei classified as cluster) were correctly segmented. Therefore, the classification error was 2.94%, all of which were, as intended, false negatives.

4.2 Validation

4.2.1. Detection of immunopositive cells

We found 2259 (V1) and 2467 (V2) objects in the two BAL validation samples, of which 49 (V1) and 99 (V2) were immunopositive tumor cells. The analysis found also 26 (V1) and 50 (V2) objects, corresponding mainly to macrophages, immunonegative cells and organic debris. Therefore, all cancer cells sprinkled in the BAL samples were satisfactorily found. Thus, the classification errors were 1.15% (V1) and 2% (V2). All errors were, as intended, false positive (FP) results.

4.2.2. Classification of nuclei vs. clusters of nuclei

The linear SVM classifier was presented with all the detected objects extracted from BAL validation samples V1 and V2. All clusters of cancer cells were properly classified as such. Seventeen percent (17.3%) of isolated cancer cells were wrongly classified as clusters, due to the high demand to avoid false positive results imposed on the classifier. These errors can be easily detected and dealt with through final visual analysis of the FISH enumeration results.

4.2.3. FISH segmentation

All objects -isolated or clustered- were imaged in 3D and analyzed to detect FISH signals in all four channels. The enumeration results were then presented to the user for validation. It is important to note that from the total number of objects existing in the samples -2259 in V1 and 2467 in V2- the cytopathologist only had to review the enumeration results in 75 (V1) and 149 (V2) objects. This -making use of our graphical user interface- can be easily done at the computer in less than 30 minutes, compared to the approximately 6 hours that would take visually analyzing both samples following the standard existing protocol.

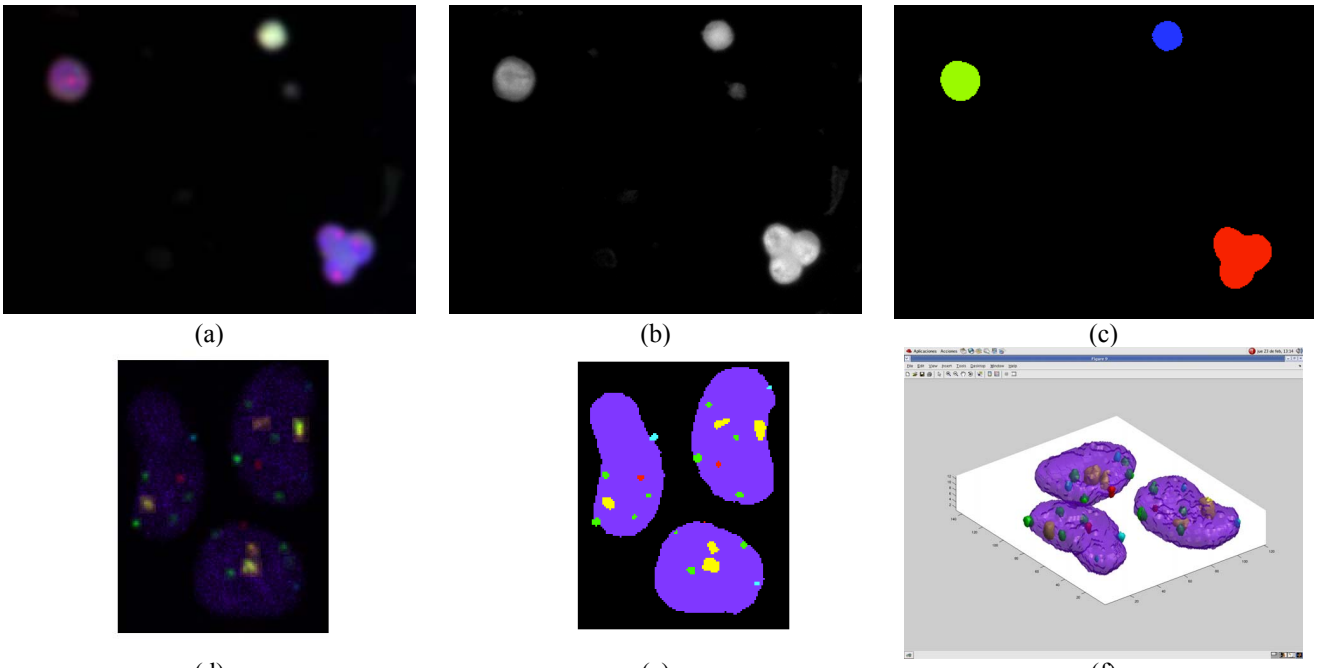

Figure 2. **Image segmentation process.** (a) Original RGB image created from the SpectrumRed (R), SpectrumAqua (G) and Alexa350 (B) fluorescent channels. (b) Immunofluorescence channel showing possible confusion between immunopositive cells (bluish in a) and macrophages (white in a). (c) Binary mask obtained after thresholding the immunofluorescence channel and classifying the segmentation results. Macrophages (blue) are eliminated after converting to HSI and thresholding the H and S values. Then the SVM classifies objects as nuclei (green) or clusters of nuclei (red) using morphological information. (d) Representative slice from a Z pseudo-colored stack that contains the counterstaining and 4 FISH DNA probes. (e) Segmentation of the nuclei (violet) and DNA probes shown in d. (f) 3D reconstruction of the entire imaged shown in d.

5. CONCLUSIONS

We have presented the hardware and software specification of an automated microscopy system to be applied to the unsupervised detection of rare cancer cells in minimal samples of lung cancer. The design of the hardware allows seamless integration of image acquisition, analysis and storage. The analysis protocol is detects immunopositive tumor cells in 2D and calculates the copy number of four DNA sequences to determine the genomic integrity of the cells, thus confirming or rejecting their tumoral origin. This task, extremely time consuming and error prone when done manually at the microscope, is in our hands free from false negative results. The moderate number false positive results - misclassified macrophages, negative epithelial or oropharingeal cells- can be easily eliminated from the analysis by looking at their genomic content, since macrophages do not hybridize any of the DNA probes and normal epithelial and oropharingeal cells have normal pattern of FISH signals. Therefore, after the automated computer work, the user is left with the simple task of reviewing copy number results of a relatively low number of cells.

6. ACKNOWLEDGEMENTS

AMC and COS hold Ramon y Cajal fellowships of the Spanish Ministry of Science and Technology. COS is currently supported by the Spanish Ministry of Science and Education (MCYT TEC2005–04732), the EU Marie Curie Program (MIRG–CT–2005–028342). This work was supported by the UTE-CIMA project.

7. REFERENCES

[1] K. Weber-Matthiesen, et al Leukemia, vol. 7, pp. 646-9, 1993.

[2] D. M. Beazley, "SWIG: an easy to use tool for integrating scripting languages with C and C," Proceedings of the 4th Conference on USENIX Tcl/Tk Workshop, 1996, vol. 4, pp. 15-15.

[3] C. Ortiz de Solórzano et al. Journal of Microscopy 193(3):212-226, 1999.

[4] V. Vapnik, “The support vector method of function estimation” in J.A.K. Suykens and J. Vandewalle (Eds) Nonlinear Modeling: Advanced Black-Box Techniques, Kluwer Academic Publishers, Boston, pp. 55–85, 1998.